DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

'NEW POLITICAL HISTORY:' SOME STATISTICAL QUESTIONS ANSWERED

J. Morgan Kousser
Allan J. Lichtman

HUMANITIES WORKING PAPER 72

February 1982

SUMMARY

In the Spring, 1981, number of Social Science History, William G. Shade defended the "ethnocultural" historians' conclusions and offered some general methodological advice to political historians. Examining his five-part argument point by point, we find his analysis of the issues muddled, his "tests" of the robustness of the ethnoculturalists' results misconceived and inconclusive, and his specific prescriptions for future historians misleading. After attempting to clarify the arguments and propose some useful guidelines, we conclude that Shade's article underlines (once again) the necessity for more intensive statistical training for social scientific historians.

`NEW POLITICAL HISTORY:'

SOME STATISTICAL QUESTIONS ANSWERED

by
J. Morgan Kousser
Allan J. Lichtman

William G. Shade's "New Political History: Some Statistical Questions Raised," has two sometimes conflicting purposes: first, to remind historians to "think statistically" and to "give more self-conscious attention to the details and logic of research design," and, second, to defend such "ethnocultural" historians as Ronald P. Formisano and Paul Kleppner against published criticisms. Too often confusing the former with the latter aim, Shade attains neither. An apology for flaws in the work of these early quantifiers interspersed with a piecemeal and misleading discussion of methodological issues, his paper is further compromised by distortions of other scholars' work and neglect of relevant literature published since 1974. Shade's list of 66 references includes only two post-1974 citations, neither of which he consulted for methodological guidance.

With Shade's two major prescriptions -- plan research carefully and use genuinely multivariate methods -- we have no quarrel. Often breached in practice, these familiar commandments can never be repeated too many times. To his lack of conceptual rigor, to his defense of previous work on the basis of a series of either

meaningless or misleading "tests," and to several of his methodological dicta, we do take exception. Thus the necessity for this note.[1]

Embedded in Shade's essay is a kind of statistical pluralism that beckons historians to choose freely from among such techniques as multiple regression, bivariate correlation using either parametric or nonparametric measures, and homogeneous area analysis.[2] Yet he offers no guidelines for the appropriate use and interpretation of these procedures. Consider his discussion of the first of his five "questions" for analysis -- "cross-level inference."

To put the discussion of cross-level inference into the proper context, let us first briefly review the "ecological fallacy" and some attempts to escape it. The ecological fallacy consists in naively inferring individual behavior from aggregate data. As Robinson (1950) showed, correlations at the two levels may differ. Benson (1961) and others responded by concentrating on those aggregate units which were more or less homogeneous in what they presumed were the vote-relevant individual traits. There are six problems with basing generalizations about whole states or regions on data from homogeneous areas. The technique excludes information from the much more numerous non-homogeneous areas, thereby restricting the variation in both independent and dependent variables and decreasing the reliability of estimates of voter behavior. It assumes that individuals residing in homogeneous units have the same mean scores on other voter-relevant variables as individuals living in heterogeneous units. The technique

ignores "contextual effects," binding investigators to the supposition that people who live in the same areas vote the same way whether their peers are similar or dissimilar to them in demographic traits or voting proclivities. Investigators using homogeneous area analysis must implicitly treat all relationships between socioeconomic characteristics as linear. Moreover, despite heroic efforts, they can usually obtain data on only a few of the possibly relevant facts about voters. Finally, the scrutiny of homogeneous areas is not compatible with the multivariate methods that Shade recommends, since it is virtually impossible to isolate units that are simultaneously homogeneous on several variables of interest. (See Kousser, 1976; Langbein and Lichtman, 1978; and Lichtman and Langbein, 1978, for more extended discussions.)

Shade's only prescription for those still determined to analyze homogeneous units is to "choose homogeneous areas in a more systematic fashion in line with acceptable sampling procedures whenever the data permit." Unfortunately, no matter how circumspect the historian's sampling scheme, it cannot surmount biases inherent in the method itself.

A more sophisticated method for bypassing the fallacy is "ecological regression," first used in the historical literature by Alexander et al. (1966) and first fully explained to historians in papers by Jones (1972), Kousser (1973), and Lichtman (1974).[3] In these articles, the authors counseled historians to use multivariate, as well as, in appropriate cases, bivariate regression (not always bivariate regression, as Shade contends), and outlined tests for the

linearity of the relationships and procedures for dealing with apparent violations of the assumption crucial to both the ecological regression and homogeneous areas techniques -- that deviations from the predicated form of the relationship were random rather than systematic.

As subsequent work has pointed out more clearly than did the early articles, a properly specified equation -- that is, one which includes all the independent variables which actually influenced the dependent variable, and which captures the proper form of the relationships (e.g., linear, log-linear, quadratic, interactive, etc.) -- will, in most cases, correctly describe individual behavior.[4] (Hanushek et al., 1974; Lichtman and Langbein, 1978.) As even the earliest articles argued, moreover, individual behavior is inferred more reliably from aggregate data through use of unstandardized regression coefficients rather than normalized measures such as correlation coefficients or standardized regression measures (beta weights). Normalized measures are uniquely subject to distortions arising from changes in the standard deviations of variables produced by the aggregation process.

Shade's own discussion of ecological inference is especially confusing. Not only does he ignore distinctions between regression and correlation measures, but he misleads readers by repeating a rule of thumb offered by Dollar and Jensen, 1971, p. 101. To simplify matters, Dollar and Jensen advised prospective users that the Pearsonian ecological correlation between two variables had to be "at least $\pm$ .7" to justify putting much confidence in ecological

regression estimates. In fact, a high correlation is neither a necessary nor a sufficient condition for an unbiased ordinary least squares regression estimate. Such an estimate will be unbiased if, and only if, there are no variables excluded from the equation which would have had non-zero coefficients had they been included, and if the form of the relationships is correctly specified. A high correlation may be a sign of a well specified equation, but it is not necessary. Suppose voting were really random across all possible groups — i.e., that an equal percentage of persons in every group voted for a certain party. Then both the correlations and the slopes of the regression lines for the percentages in each group, class, or whatever and the vote would be zero (assuming no aggregation bias), but the estimates of how each group voted would be perfectly correct.

In a related error, Shade misinterprets Goodman, 1959, by stating that the "no excluded correlated variables" assumption just discussed amounts to an assumption that "The patterns of settlement . . . produced a random distribution of ethnic groups." (Shade, 1981, p. 1976.) What is at issue in this assumption is not a random distribution across counties or townships, but random deviations of the points representing those units from the true regression line or surface in the population of interest.

Comparisons of the results of regression estimates with actual votes from poll book, survey, and other individual level data have uniformly found the estimates from carefully chosen equations based on aggregate data to be quite close to the actual behavior patterns of individuals in a fairly wide variety of cases. (Irwin, 1967; Stokes,

1969; Kousser, 1973; Bourke and DeBats, 1980.) These tests increase one's confidence in the usefulness of the technique.

Shade proposes to "test" for the "pitfalls of ecological regression" by comparing a bivariate regression estimate of the proportion of Germans who voted Democratic in 1851 in Pennsylvania, calculated from data from all the state's counties, with two other estimates: the first derived from an analysis of votes in homogeneous minor civil divisions in one county, and the second, from an ecological regression of the percentage of votes for each party on the percentage of Germans by townships in that county. No individual-level data on German voting behavior for this election seems to have survived. Finding that the countywide regression and the homogeneous area estimates are equal to one another and differ from the statewide estimates by 20%, he concludes that the results of the ethnocultural historians for a number of states over a great many elections would probably not have differed had they used ecological regression, and that in the future, historians should feel entirely free to use either homogeneous areas or regression analysis "depending on the nature of the available data and the questions being asked." (Shade, 1981, p. 177.) He offers no further guidance on which questions are appropriate for each technique.

But Shade's test is not a proper test for seven reasons. First, the equation at the statewide level is misspecified unless all other determinants of voting behavior (for example, religion or class) were uncorrelated with ethnicity. His is one regression estimate of the behavior of individual German voters, but almost certainly not the

best one which could have been made.[5] Second, while a comparison of a statewide estimate based on data from all the counties against an estimate using all the civil divisions or townships for the whole state -- not just for one county -- might be an appropriate way to test for possible bias induced by the combination of smaller areas into counties, even this comparison would not uncover possible biases introduced by combining voters into townships.[6] Shade simply has no individual-level data, and therefore has no information appropriate for a test of how well ecological regression predicts individual-level behavior.

Third, the estimates from one county are no doubt based on a small number of cases (Shade never says how many) and therefore, even if unbiased, probably have larger standard errors than the statewide estimates (he did not include standard errors or "t" tests for either regression in any version of his paper). Consequently, the Schuylkill county estimates are almost certainly less precise, in a statistical sense, than the statewide estimates. Fourth, even if unbiased and precise, the estimates might have come from a county which for some reason deviated from the statewide relationship. Shade never makes clear how much Schuylkill county deviated from the linear regression line. If Schuylkill were a randomly deviant case, the statewide estimates might still be correct, but the Schuylkill and statewide results would differ. Fifth, some of Shade's comments indicate that Schuylkill was a case which systematically deviated from the statewide linear relationship. The "premier mining area in eastern Pennsylvania" and later focal point for the Molly Maguires, it had a

German plurality, a Welsh and English minority, and a "rapidly growing group of Catholic Irish." (Shade, 1981, p. 175.) This ethnic stew might well have raised group conflict in voting to levels above that in the average county, thereby increasing the correlation between ethnicity and the vote. It is perfectly possible that 70% of the county's Germans, but only 50% of the state's, voted Democratic. And if people in other counties which were similar to Schuylkill behaved in this manner, that fact would indicate that the statewide estimate should be based on a nonlinear, not a linear functional form. Further, if miners voted differently than non-miners, a properly specified equation for the state should include a measure of mining activity. Shade takes neither of these factors into account in his statewide estimate.

Sixth, even if ecological regression "failed" a proper test, that fact would not validate the homogeneous areas method. Shade does not compare an estimate from homogeneous areas to individual statewide or even county-wide data. Seventh, even if he had made a proper test, that would not invalidate regression or validate homogeneous areas absolutely unless he were willing to argue that the same results held everywhere throughout all time, or at least in the nineteenth century in areas that the ethnoculturalists have studied.

In sum, Shade has used an almost assuredly misspecified equation to make a statewide estimate, compared it not with individual, but with aggregate estimates of voting behavior from one probably deviant county, and then, without answering criticisms pointedly made available to him five years before he published, used

this one-county, one-election "test" to validate the use of estimates from homogeneous areas in elections over most of the north for much of the nineteenth century. This parody of statistical procedure should neither comfort Shade's friends nor guide future historians.

Nor should Shade's discussion, in the second section of his paper, on levels of measurement. In an attempt to devise post-hoc alibis for his ethnoculturalist compatriots, Shade stuffs and demolishes a straw man, treats published criticisms very selectively, consequently purveying misleading advice, sets up another meaningless "test," and then misreads his own results.

The "straw man" is that critics have charged that treating interval (numeric) data as ordinal (ranked) biased the results of the "new political historians." So far as we know, this charge was never made, and it is certainly not present in the paper (Kousser, 1976, pp. 9-10) which Shade cites. (Shade, 1981, p. 178.) Of the three points which Kousser did make, Shade more or less admits the first -- that information on the exact extent of differences in variables is squandered when analysts use rank-order correlations -- but ignores the other two: that a resort to rank-order measures makes testing for nonlinear effects (e.g., polynomial, multiplicative interaction, logarithmic) impossible and that it forces one to rely on less powerful significance tests to assess whether two or more variables are associated or not. A fourth point, stressed throughout, but not explicitly stated during Kousser's discussion of nonparametrics, was that the use of Spearman's Rho or Kendall's Tau

precludes historians from developing multivariate models.

Shade's "test" is to calculate Spearman's Rho and bivariate Pearson's "r" coefficients for county-level data on the relationships between the vote in a "Maine Law" (temperance) referendum in Pennsylvania in 1854 and ten social or economic variables. Ignoring criticisms of the use of zero-order, ecological correlations offered, for instance, in the paragraph preceding the discussion of nonparametrics in Kousser's 1976 article, criticisms which Shade notes and admits earlier in his own paper, Shade is content to eyeball two sets of flawed measures and conclude summarily that they both lead to similar conclusions.

But even accepting the validity of his test for purposes of argument, his conclusion does not follow. Careful examination of the two sets of bivariate correlation coefficients reveals potentially important divergences in historical interpretation, not the "modest differences" which Shade claims. Following his own procedure of rating the importance of variables according to their capacity to predict variation in the dependent variable, we note that the pattern of his Spearman's Rho coefficients supports a religious interpretation of temperance voting, whereas that of his Pearson's "r" coefficients suggests an ethnic interpretation. His column of Rhos indicate that the proportion of Presbyterians in a county was the most important determinant of temperance voting, followed by the proportion of Methodists. But in his column of Pearson "r"'s, the proportion of Pennsylvania Dutch and the proportion of English in a county were the most crucial factors. The proportion of Presbyterians fell to fourth

place and the proportion of Methodists, to eighth place among the ten coefficients.

We begin the discussion of Shade's third topic, significance tests, by first setting out our position without reference to his. Two central priorities of the "new" historians have been to give numerical precision to such verbal expressions as "more," "less," and "most," and to assure that quantitative analyses are as reliable as possible. Statistical methodology offers historians both a ready stock of numerical measures and a means for reasoning formally about error. Significance tests and related procedures are designed to avoid the confounding of substantive results with artifacts produced by random error — a form of error that, with equal probability, generates positive or negative discrepancies between measured and actual results. For example, statements that more Whigs than Democrats voted for temperance, or that a higher percentage did, or that ten percent more did are not very interesting unless we can first reject the hypothesis that such observed differences reflect random error rather than actual behavior. Although a significance test cannot, of course, establish the substantive importance of a result, it can, as David Gold, one of Shade's own sources, has noted, provide a useful check against drawing conclusions from "an observed association [that] could be generated in a given set of data by a random process." (Gold 1963, p. 46.)

In Shade's view, however, the ethnoculturalists should be exempt from the usual canons of research procedure since "they

generally did not deal with systematically drawn samples, but with `total populations.'"[7] Yet numerous sources of random error can afflict measures computed for "total populations" as well as for samples of data. Consider four examples in which significance tests can be useful correctives to unwarranted inferences, even when studying "total populations." First, random measurement error may arise either from the original collection of information or the historian's own processing of data. Second, the historian may be able to measure only proxies for the variables that truly are of interest. For example, an investigator might use occupation or education as a proxy for social standing or church seats as a proxy for religious affiliations. Measures computed from such proxy variables may differ both randomly and systematically from the true values for the variables which are really of interest. Third, whenever historians infer conclusions about individuals from data collected for aggregate units, they are, in effect, engaged in sampling. In this case, the units studied by the investigator represent a sample of cases drawn from the total population of individuals according to the process by which they were arranged into geographical subdivisions. As in any sampling process, such grouping can generate both random and systematic error. Finally, the "total population" studied by the historian may only be a subset of the population whose behavior he actually seeks to explore. For instance, the historian may have data for every legislator in a given statehouse for a roll call on a temperance law. But he may truly be interested in whether all Whigs and Democrats, voters as well as legislators, held different

attitutdes on temperance. The historian might also consider the roll call votes on specific proposals as samples of the legislators' attitudes on the general subject of liquor control. In these cases, we can use significance tests to determine whether this "sample" of the population's underlying attitudes indicates that partisan views on the subject of drinking, which is the "population" really of interest, diverged or not.

Although Shade recognizes the force of this fourth argument, he sidesteps it by asserting that "the significance test is a useful tool" only "if one is dealing with samples that have known probabilities," whereas "each of these authors conceived of his total population as a nonprobability sample of a larger universe." With this argument Shade inadvertently concedes that ethnocultural scholarship suffers from a more serious problem than random error. For if these historians were generalizing from nonprobability samples, their results may have been marred by systematic biases that yield results which are skewed in a particular direction and which therefore cannot be detected by the usual significance tests. The fact that historians should, as Shade admonishes, "consider the associations between their samples and the universes to which they wish to generalize," but that the ethnoculturalists, according to Shade, did not do so, reduces these historians' credibility even further.

Yet that observation would not free them from using significance tests. It merely implies that they and other historians should try to reason about the biases involved in their data and if possible redefine the significance tests accordingly. If their

samples were skewed toward finding, for example, more cohesive German Lutheran and Pietist political behavior than average in the state or region, they might greatly increase the required significance level on a test for differences in the two groups' voting records in order to take the sampling bias into account. Since homogeneous areas probably exhibited more uniform voting patterns than heterogeneous ones, it would be reasonable to infer that two groups' behavior differed in the total population only if one found a very great divergence -- more than one would expect using a conventional level of significance -- in voting returns drawn exclusively from such areas.

Finally, Shade raises questions about the choice of particular levels of statistical significance, in the process confusing scientific convention with "subjectivity." It is surely true, as is invariably noted in the first few weeks of any introductory statistics course, that there is nothing sacred about 0.05, 0.01, or, for that matter, 0.75. The reason for using "low" significance levels is that, otherwise, it would be too easy to reject null hypotheses, and science would become violently unstable, as every new study overturned a previous one. But while any particular level is arbitrary, it at least provides a precise, and in that sense, objective decision rule for accepting or rejecting a finding, and one which may well be widely agreed upon by scholars of diverse disciplines and interests.[8] Thus, for someone to publish coefficients significant at the 0.25 level would raise numerous eyebrows.

Consider Table 1, which is based on an 1840 roll call in the lower house of the Michigan legislature on banning railroad trains

from running on Sunday. (Formisano, 1971, pp. 123-24.) To decide whether this difference in party behavior, taken by Formisano as indicative of the "central tendencies of the parties" on such issues, is sufficiently large for reliable inference, we computed a Chi-Square statistic. The Chi-Square value is 0.518, which is significant only at the 0.47 level. While statisticians often urge analysts to publish the actual significance levels of their parameters, and not just to denote which of them passes the 0.05 or 0.01 barriers, few social scientists would feel comfortable printing "0.47" as an attained significance level at the bottom of a table, and few readers would put much credence in conclusions based on the contention that despite a tiny Chi-Square, the relationship in question was actually strong.[9] While significance levels are only conventions, they are useful ones.

TABLE 1: 1840 MICHIGAN HOUSE SABBATARIAN VOTE

Attitude On Banning Sunday Travel	Party	
	Whig	Democrat
For	19 (66%)	7 (54%)
Against	10 (34%)	6 (46%)
TOTAL	29(100%)	13(100%)

In sum, although Shade's warnings about the unthinking use of significance tests and the confusion of statistical with substantive importance are worth heeding, he confuses on this topic more than he guides. If he is interpreted as encouraging historians to abjure tests of significance altogether, which is not an unreasonable reading of some of his remarks, the result will be a move away from instead of towards proper methodological practice.

Nor is Shade's discussion of "synoptic measures" a positive aid to understanding. Suppose a historian believes that the Democratic vote (D) in some election depended on the number of Irish (I), Germans (G), and Episcopalians (E) in each county in some state. Then he might formulate and test his hypothesis in the usual multiple regression manner as:

$$(1) \quad D = \beta_0 + \beta_1 I + \beta_2 G + \beta_3 E + u,$$

where the β's are coefficients to be estimated and u is an error term. But since many analysts are interested less in the actual vote than in the percentage the party received, they would find it more natural to express their hypotheses as, e.g., the greater the <u>proportion</u> of Irish, etc., the greater the <u>proportion</u> of Democrats. This second hypothesis would take the form:

$$(2) \quad D / P = \beta_0 + \beta_1 I / P + \beta_2 G / P + \beta_3 E / P + u,$$

where P is the number in the eligible voting population or perhaps the number who actually voted.

The reader will note that P appears as a denominator for the variables on both sides of the equation. The dependent variable is,

therefore, being regressed on independent variables which are, by definition, partly functions of itself. Will this fact artifactually inflate the estimates of the regression parameters (the β's)? The short answer is that it depends on which hypothesis the historian believes -- that given in equation (1) or that encompassed in (2). If the predicated relationship is between proportions, then the coefficients will not be higher than they "should" be; on the other hand, if the theory properly relates numbers of people, and the error term meets the usual assumptions, then there is no particular reason to normalize by population or by any other quantity.[10]

Recognizing this point, Shade argues that population size may still be important as a proxy for urban/rural differences in voting; urges historians to "control for size statistically" by introducing population as an independent variable in equations such as (1); performs another "test" to see whether the ethnoculturalists' failure to introduce such a control distorted their findings; concludes that it did not; and closes the section by repeating his homily about the necessity for "carefully formulated hypotheses." (Shade, 1981, pp. 184-86.) His discussion is flawed on several counts.

First, if an historian thinks that an urban/rural split was an important determinant of voting -- a proposition which seems not to mesh with the ethnocultural thesis -- the percentage living in cities or towns would appear to be the natural proxy to choose. If one used population, instead, for data drawn from the 1850s, for example, it might well be that a geographically large and densely populated rural county would be judged, by population size, more "urban" than a

geographically smaller county where nearly everyone lived in a town.

Second, Shade errs in suggesting that the proper adjustment for differences in voter turnout is to "control for [population] size statistically." Such controls index only the influence of differences in the number of potential voters, not in voter turnout. Investigators can take turnout into account by measuring both independent and dependent variables using the potential voting population rather than the vote cast as the denominator for percentages. Analysis of such variables reveals the relative support given candidates and parties by groups within the total potential electorate. Historians can also gain insight into turnout effects by using the proportion of voters in the potential electorate as a dependent variable and by employing regression techniques for measuring transition probabilities between voting and nonvoting.

Third, Shade uses yet another misleading test to exonerate the ethnoculturalists from responsibility for controlling for population size, as he himself recommends to future investigators. In his "test," Shade estimates a series of equations such as:

$$
\begin{aligned}
&T = \beta_0 + \beta_1 I + \beta_2 P + u, \\
(3)\quad &T = \beta_0 + \beta_1 G + \beta_2 P + u, \\
&T = \beta_0 + \beta_1 M + \beta_2 P + u,
\end{aligned}
$$

where T stands for the vote in the 1854 Maine Law referendum in Pennsylvania, M, for Methodists, all the rest of the variables are as defined earlier and the data is aggregated at the county level.[11] He then compares the partial correlation coefficients, not the partial

regression coefficients, for the variables I, G, M, and so on (but not for P) with zero-order correlation coefficients computed from equations of the form:

$$
\begin{aligned}
&T / P = \beta_0 + \beta_1 I / P + u, \\
(4)\quad &T / P = \beta_0 + \beta_1 G / P + u, \\
&T / P = \beta_0 + \beta_1 M / P + u.
\end{aligned}
$$

Since these two sets of correlation coefficients are "roughly" similar, he again exonerates the ethnoculturalists, who, when they computed statewide statistics, used equations like (4), and not, as he favors, equations like (3).

As with his previous "tests," this one is seriously deficient. Every bivariate or trivariate equation he uses is surely misspecified and the parameters are therefore biased. A comparison of two sets of biased coefficients is hardly conclusive evidence that one of them is not biased. Even ignoring bias, correlation coefficients are, for reasons detailed in our 1973 and 1974 articles and reiterated above, inferior statistics for aggregate data analysis. Furthermore, by comparing equations like (4) to equations like (3), Shade is, in effect, contrasting two rather different, though related, theories, one based on votes and the number in each group, and one based on proportions.[12] It would therefore be a bit difficult to know what to expect from such a comparison or what to make of the results, even if a meaningful test had been performed.

Shade's results are also difficult to interpret since his zero-order correlations are reported only for "Pro-temperance"

(Table 1) voting and his partial correlations (controlling for population) for both "Pro-Temperance" and "Anti-Temperance" voting (Table 2). The differences are potentially important. The partial correlation for "Farm Value" voting is -.0785 for Pro-Temperance voting, but .5007 for Anti-Temperance voting. Unfortunately, we are not supplied the information necessary either for explaining this difference or for determining whether it is also present in the zero-order coefficient. Nonetheless, the ethnoculturalists once again emerge from Pennsylvania in fine shape as Shade emphasizes "that the basic relationships remain the same" whether or not population is controlled.

It is not only his "test," however, that is inadequate here: his whole section on synoptic measures misleads. The chief problem with leaving out population is that doing so often cases "heteroskedasticity," or unequal variances of the error terms for each unit or observation.[13] While heteroskedasticity does not produce biased parameter estimates in ordinary least-squares regression, it does increase the variance of the estimates and it invalidates the usual significance tests. The standard solution, as outlined more extensively, e.g., in Kousser, 1980, is to weight each variable in equations such as (2) by $\sqrt{P_i}$. In this, as in other sections of his article, then, Shade's analysis raises important points without clarifying them, his "test" is deceptive, and his suggestions for future work are counterproductive.

Shade's fifth and final "question" is whether the

ethnoculturalists' use of bivariate correlation, rather than what Shade implicitly admits in this section of his paper is the superior technique of multiple regression, might have led them to adopt an ethnic or ethnoreligious, rather than an economic interpretation of American politics. He "tests" this possibility by regressing the county-level election returns in the 1854 Pennsylvania temperance referendum, in a stepwise fashion, on one ethnic (percent English), one religious (percent Pennsylvania Dutch), and one economic (farm value) variables and determining how much additional variance in the wet and dry percentages each type of variable explains.

With its exclusive focus on a temperance referendum, Shade's procedure cannot determine whether ethnocultural divisions pervaded Pennsylvania's partisan contests in the 1850s, but can only shred a straw man of his own creation. That liquor laws divided nineteenth century ethnic groups was no discovery of Benson, Hays, et al., nor did they claim such originality. Historians have long known that Germans like their beer, Irish, their whiskey, New Englanders, their cold water. It is precisely in the response to temperance laws that one would expect ethnic and religious variables to weigh in most heavily as determinants of voter choice.

Shade's operational measures and specific procedures also raise serious questions about the logic and execution of his "test." "Farm value" has the wrong sign in one part of his Table 3, no doubt because of a misprint. Shade's index of "Pennsylvania Dutch," based on church seats, is oddly almost perfectly positively correlated with his "German Orthodox" variable and nearly perfectly negatively

correlated with his percent "English." (Shade, 1981, Tables 1 and 3.) After all his strictures about "holding population constant statistically" earlier in his paper, he does not include population in his multiple regression. Several variables with relatively high zero-order correlations with temperance in Table 1 are not entered into the multiple regression in Table 3. The statistician's model does not demand, as he implies (p. 191), orthogonal independent variables in multiple regression -- indeed, if all independent variables were mutually orthogonal, bivariate methods would suffice.

In any case, the assessment of the influence of different factors through stepwise regression and the "additional variance explained" (incremental increase in R^2) criterion is a misleading procedure that attributes all the variance mutually explained by correlated variables to whichever variable is first entered in the equation. For example, any part of the correlation with "dry" sentiment explained by both farm values and the percentage English is chalked up entirely to the English. The economic variable (or any variable entered later than the percentage English) gets a "chance" to explain only the residual variation in the dependent variable left after the English variable has explained everything it could. Whatever its actual importance, the contribution of R^2 of the nth variable entered in a regression equation cannot be greater than $1 - R^2_{n-1}$ where R^2_{n-1} is the value of R^2 attained prior to the inclusion of the nth variable. Since Shade's technique treats variables, in this sense, asymmetrically, it is a deficient method for comparing the importance or the contributions of different variables to explaining

voting behavior.

Aside from problems of asymmetrical measurement, an exclusive focus on explained variance is misguided for cases of cross-level inference, such as Shade's example. Since the most spirited and sophisticated defender of such a focus is John Hammond, and since Shade provides no such defense, we will consider Hammond's arguments. In two thoughtful articles and a book, Hammond (1973, 1979a, 1979b) advocates the use of standardized regression coefficients for rating the importance of variables according to their contribution to explained variance. In particular, he recommends using beta weights for aggregate data in which the variables are measured with error that has a particular (multiplicative) structure or in which the variables are assumed to be arbitrarily scaled indicators of underlying attitudes (1979a, pp. 478-83). He also points out that beta weights (which are identical to Pearsonian correlation coefficients in the bivariate case) allow a comparison of the effects of variables which do not have a common scale, such as most measures of ethnicity and economic welfare.

Although Hammond's points are well-argued, on balance, we reject his advice for six reasons. First, as Hammond himself admits (1979a, p. 485), standardized coefficients for aggregate data may be biased estimates of individual-level relationships in cases where unstandardized coefficients are unbiased. The value of beta weights, like other normalized measures, is a function both of individual-level relations between independent and dependent variables and of differences produced strictly by the grouping process in the relative

variance of competing independent variables.[14] Indeed, such differences in relative variance are precisely what would be expected in virtually every case of interest to historians, as groups are almost invariably distributed differently across geographical units. Second, as he shows in his 1973 article, if one group is more geographically concentrated than another, the aggregate estimate of the individual-level standardized coefficient will be more inflated for the more segregated group, although the unstandardized estimates may well be unbiased for each. Thus, a comparison of the relative magnitudes of the two standardized coefficients will exaggerate the importance in explaining the dependent variable of the more segregated, relative to the less concentrated group, in many cases in which the same comparison for unstandardized coefficients will not. Third, since aggregation processes are likely to produce different changes in relative variation, the values of beta weights will depend on the particular set of units chosen for analysis. This means that even for properly specified models, ecological inference will be unstable as the analysis shifts from one level of aggregation to another.

Fourth, for multivariate equations, the standardized regression coefficients have no natural interpretation. In particular, they are not truly measures of the percentages of variance explained by independent variables. As theorist Hubert Blalock notes, "The partial correlation is a measure of the amount of variation explained by one independent variable . . . The beta weights, on the other hand, indicate how much change in the dependent variable is

produced by a standardized change in one of the independent variables." (Blalock, 1972, p. 453, italics his.) Fifth, we see no reason to believe that, in general, measurement error for aggregate-level variables is multiplicative. Indeed, most of the examples of multiplicative error cited by Hammond (such as using the percentage of residents born in Scandinavia to infer the behavior of all generations of Scandinavian-Americans) can be conceptualized as specification error and treated accordingly. Sixth, as a sociologist, Hammond may wish to ignore the specifics of the historical situation, such as how referendum questions were posed or differences in the demographic specifics of various groups -- their age and sex composition, recency of migration, voting turnout, etc. (Hammond, 1979a, pp. 483-84). As historians, we believe all these specifics are potentially important, and we want to avoid using a method which would make it easy to ignore such factors.[15]

We know Bill Shade to be an honest and intelligent person and a careful and productive scholar. What, then, accounts for his publication, after a lengthy and meticulous reviewing process (in which we were not formally involved) of such an article, and what does it imply for the historical profession? We surmise that a desire to preserve a paradigm to which he is attached and to defend a group of scholars with whom he has been personally associated explains Shade's willingness to propose seriously deficient ad hoc "tests" and to reject criticisms of them. Thomas Kuhn (1970) has made us all aware that such strategems are a normal part of science as well as a part of

"normal science." Perhaps Shade's reviewers shared his motives. Perhaps, also, the impulse to seek some detached judge in such cases is as futile as the search for an Archimedian fulcrum. Even so, we believe that more statistically sophisticated referees would not have approved the piece, even if they were sympathetic to Shade's point of view. While we agree with Shade's call for better and more self-conscious research designs and for the adoption of genuinely multivariate methods, we think that the misconceptions of his article underline to an even greater extent the necessity for much more thorough statistical training for quantitative social scientific historians.

Footnotes

1. It may be of interest to the reader to learn of our previous connection with Shade's article. Originally delivered at a 1976 session of the Social Science History Association convention which Kousser organized and chaired and at which Lichtman and Irwin gave a paper, Shade's piece was revised, after our oral criticisms both before and after the panel. In late January, 1977, Shade sent Kousser a copy of the revised version, which Kousser critiqued in an April, 1977, letter to Shade. Shade's present paper, which has been only slightly revised from the January, 1977 version, takes no account of our earlier criticisms and includes only one reference to work published since 1976, despite the publication between then and now of numerous relevant works. In fact, citations to the Hanushek et al. (1974) and Lichtman-Irwin papers, which were published in Political Methodology in 1974 and in SSH in 1978, were even deleted between the 1977 and the most recent version. The principal criticisms we offer now should come as no surprise to Shade, since we have made them to him before.

2. While multiple regression, on which Shade concentrates, is undoubtedly useful, he might also have mentioned more advanced techniques, which have recently been introduced into the historical literature, such as logit and probit analysis.

On these, see Knoke and Burke (1980), Kousser (1980), and Goldin (1981).

3. Shade distorts the work of Alexander et al. and McCrary et al. (1978) when he says that the contrasts in their findings, both based on ecological regression, "can only be resolved by a methodological `leap of faith'." (Shade, 1981, p. 173.) While the McCrary results were based on multiple regression analyses for all Alabama counties, Alexander's rested on regressions with data drawn from non-random surviving beat (the Southern equivalent of township) returns in only fifteen of the state's counties. It is hardly surprising that analyses based on different universes of data led to different results.

4. It is misleading to put as much stress on the assumption of constant behavior across geographical units as, for example, Vinovskis (1980) does. That assumption is relatively easy to test and correct for in practice. The much graver difficulties in deciding whether individual-level inferences are right or not arise from the possibility of specification error and aggregation bias, on which see Lichtman and Langbein, 1978.

5. In addition to the criticisms offered in the text, it must be noted that even were Shade convincing, his analysis here would support not an ethnocultural, or ethnoreligious, but merely a more primitive ethnic hypothesis.

6. For an interesting test of the differences between county- and

civil-division-level regression estimates for Iowa in 1924, see Waterhouse, forthcoming.

7. Few if any of the controversialists in Morrison and Henkel (1970), Shade's main source of criticisms of the use of significance tests, discuss the question of running such tests on "total populations." Shade also ignores the effective criticisms of the 1957 article by Selvin, which sparked the debate in sociology, in numerous other articles in the book, as well as the fact that the whole controversy has seemingly died away in sociology since 1970. Indeed, sociologists are increasingly turning to such techniques as log-linear modelling, which involves wholesale computations of Chi-Square values to evaluate different hypotheses. See, e.g., Goodman, 1978.

8. Of course, it is possible to imagine situations in which standard significance tests and conventional levels of certainty ought to be jettisoned. Since if one is working with a very large number of observations, chance alone will often produce apparently significant relationships between variables at the 0.10 or 0.05 levels, one ought in such cases to impose more stringent criteria for significance. If one's sample were skewed in known ways or if one had a snarply peaked "prior" belief about some outcome, a redesigned test or perhaps an unusual significance level would provide a more appropriate decision rule. Alas, historians generally operate in a world of diffuse priors and samples of unknown bias. Armed only with a rough set of hypotheses, they

are presented with a bunch of numbers whose representativeness they can determine only approximately, and they must say to themselves: "On the basis of this collection of data, how much credence should I give to this hypothesis?" As a practical matter, conventional significance tests are often the only stop this side of relativism.

9. Formisano runs no such significance test and this is the only legislative vote on "moral" issues for which he provides a party breakdown. If one adds the two Whig and four Democratic abstainers to the table, the Chi-Square rises to 3.412, which is significant at the 0.18 level.

10. Bollen and Ward's 1980 article provides a good introduction to the literature on this highly controversial subject. Because the problems of multicollinearity and misspecification, discussed below in the text, seem to us especially grave in the equation (1) form of the hypothesis, we take a somewhat different position on the use of ratio variables in this particular case than they do for the general case.

11. It is possible that the actual set of equations Shade uses is of the form $T/P = \beta_0 + \beta_1 I/P + \beta_2 P + u$. His discussion is not clear on the point.

12. If, of course, population size <u>per se</u> is an independent "contextual" influence on aggregate-level voting, then its exclusion from a regression equation will bias parameter

estimates whether theory calls for specification in ratios or raw numbers. In this special case, population should be entered as an additional control variable. If, however, equations are specified in terms of raw numbers rather than ratios, population size will be highly collinear with the numbers of people in the larger population groups, creating severe problems of multicollinearity in regression estimates.

13. There is a good discussion of hetereoskedasticity and autocorrelation, with which Shade confuses it, in Hanushek and Jackson, 1977, pp. 142-46.

14. This follows directly from the formula defining the standardized regression coefficient, which is the unstandardized regression coefficient multiplied by the ratio of the standard deviation of the independent to the dependent variable. For example, in the three variable case: Byx.z = byx.z Sx/Sy, where Byx.z is the beta weight, byx.z the unstandardized correlation coefficient, Sx the standard deviation of X, and Sy the standard deviation of Y. Even when byx.z at the aggregate level is an unbiased estimator of its individual level counterpart, Byx.z will not be an unbiased estimator of the individual level beta weight, except in the unlikely event that Sx/Sy remains unchanged after aggregation. Hammond is certainly aware of how grouping alters relative variance, but oddly concludes that such changes bias unstandardized, but not standardized coefficients (Hammond, 1979a, pp. 478-82). For more detailed discussion and empirical

examples see Langbein and Lichtman, 1978, pp. 36-38.

15. In the execution of multivariate analysis, we would stress not only the estimation of particular parameters, but also the proper form (e.g., multiplicative, linear, interactive) of the multivariate model itself. (See Broder and Lichtman, 1982.)

References

Alexander, Thomas B., P. Elmore, F. Lowery and M. Skinner (1966) "The Basis of Alabama's Two-Party System," Alabama Review, 19, 243-76.

Benson, Lee (1961) The Concept of Jacksonian Democracy: New York as a Test Case. Princeton: Princeton University Press.

Blalock, Hubert M. (1972) Social Statistics. New York: McGraw Hill.

Bollen, Kenneth A. and Sally Ward (1980) "Ratio Variables in Aggregate Data Analysis: Their Uses, Problems, and Alternatives," in Edward F. Borgatta and David J. Jackson, eds., Aggregate Data: Analysis and Interpretation. Beverly Hills, Calif.: Sage Publications: 60-79.

Bourke, Paul F. and Donald A. DeBats (1980). "Individuals and Aggregates: A Note on Historical data and Assumptions," Social Science History, 4: 229-50.

Broder, Ivy and Allan J. Lichtman (forthcoming, 1982) "Modeling the Past: A Note on the Search for Proper Form," Journal of Interdisciplinary History.

Dollar, Charles M. and Richard J. Jensen (1971) Historian's Guide to Statistics: Quantitative Analysis and Historical Research. New York: Holt, Rinehart and Winston, Inc.

Formisano, Ronald P. (1971) The Birth of Mass Political Parties: Michigan, 1827-1861. Princeton: Princeton University Press.

Gold, David (1969) "Statistical Tests and Substantive Significance," American Sociologist, 4: 42-46.

Goldin, Claudia (1981) "Family Strategies and the Family Economy in the Late 19th Century: The Role of Secondary Workers," in Theodore Hershberg, ed., Philadelphia: Work, Space, Family, and Group Experience in the 19th Century. New York: Oxford University Press: 277-310.

Goodman, Leo S. (1959) "Some Alternatives to Ecological Correlation," American Journal of Sociology, 64: 610-25.

_____ (1978) Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis. Cambridge, Mass.: Abt Books.

Hammond, John L. (1973) "Two Sources of Error in Ecological Correlations," American Sociological Review, 38: 764-77.

_____ (1979a) "New Approaches to Aggregate Electoral Data," Journal of Interdisciplinary History, 9: 473-92.

_____ (1979b) The Politics of Benevolence: Revival Religion and American Voting Behavior. Norwood, N. J.: Ablex Publishing Corporation.

Hanushek, Eric A. and John E. Jackson (1977) Statistical Methods For Social Scientists. New York: Academic Press.

Hanushek, Eric A., John E. Jackson and John F. Kain (1974) "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy," Political Methodology, 1: 89-107.

Irwin, Galen Arnold (1967) "Two Methods for Estimating Voter Transition Probabilities," Unpub. Ph.D.thesis: Florida State University.

Jones, E. Terrence (1972) "Ecological Inference and Electoral Analysis," Journal of Interdisciplinary History, 2: 249-62.

Kousser, J. Morgan (1973) "Ecological Regression and the Analysis of Past Politics," The Journal of Interdisciplinary History, 4: 237-62.

_____ (1976) "The New Political History: A Methodological Critique," Reviews in American History, 4: 1-14.

_____ (1980) "Making Separate Equal: Integration of Black and White School Funds in Kentucky," The Journal of Interdisciplinary History, 10: 399-428.

Knoke, David and Peter J. Burke (1980) Log-Linear Models. Beverly Hills, Ca.: Sage.

Kuhn, Thomas S. (1970) The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Langbein, Laura I. and Allan J. Lichtman (1978) Ecological Inference. Beverly Hills, Ca.: Sage.

Lichtman, Allan J. (1974) "Correlation, Regression, and the Ecological

Fallacy: A Critique," The Journal of Interdisciplinary History, 4: 417-33.

Lichtman, Allan J. and Laura I. Langbein (1978) "Ecological Regression Versus Homogeneous Units: A Specification Analysis," Social Science History, 2: 172-94.

Morrison, Denton E. and Ramon E. Henkel (1970) The Significance Test Controversy - A Reader. Chicago: Aldine Publishing Company.

Robinson, W. S. (1950) "Ecological Correlations and the Behavior of Individuals," American Sociological Review, 15: 351-57.

Shade, William A. (1981) "'New Political History:' Some Statistical Questions Raised," Social Science History, 5 (Spring, 1981), 171-96.

Stokes, Donald E. (1969) "Cross-Level as a Game Against Nature," in Joseph Bernd (ed.), Mathematical Applications in Political Science, Charlottesville, Va.: University of Virginia Press, 62-83.

Vinovskis, Maris A. (1980) "Problems and Opportunities in The Use of Individual and Aggregate Level Census Data," in Jerome M. Clubb and Erwin K. Scheuch, eds., Historical Social Research: The Use of Historical and Process-Produced Data. Stuttgart, Germany: Klett-Cotta, 1980, 53-70.

Waterhouse, David (forthcoming) Journal of Social History.